Behavioral/Systems/Cognitive

Determining the Neural Substrates of Goal-Directed Learning in the Human Brain

Vivian V. Valentin,[1,2] **Anthony Dickinson,**[3] **and John P. O'Doherty**[1,2]

[1]Division of Humanities and Social Sciences and [2]Computation and Neural Systems Program, California Institute of Technology, Pasadena, California 91125, and [3]Department of Experimental Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom

Instrumental conditioning is considered to involve at least two distinct learning systems: a goal-directed system that learns associations between responses and the incentive value of outcomes, and a habit system that learns associations between stimuli and responses without any link to the outcome that that response engendered. Lesion studies in rodents suggest that these two distinct components of instrumental conditioning may be mediated by anatomically distinct neural systems. The aim of the present study was to determine the neural substrates of the goal-directed component of instrumental learning in humans. Nineteen human subjects were scanned with functional magnetic resonance imaging while they learned to choose instrumental actions that were associated with the subsequent delivery of different food rewards (tomato juice, chocolate milk, and orange juice). After training, one of these foods was devalued by feeding the subject to satiety on that food. The subjects were then scanned again, while being re-exposed to the instrumental choice procedure (in extinction). We hypothesized that regions of the brain involved in goal-directed learning would show changes in their activity as a function of outcome devaluation. Our results indicate that neural activity in one brain region in particular, the orbitofrontal cortex, showed a strong modulation in its activity during selection of a devalued compared with a nondevalued action. These results suggest an important contribution of orbitofrontal cortex in guiding goal-directed instrumental choices in humans.

Key words: **decision making; habit learning; instrumental conditioning; orbitofrontal cortex; outcome devaluation; reinforcement learning; fMRI; reward**

Introduction

Instrumental conditioning involves learning to perform a particular action in response to a stimulus to obtain rewards or avoid punishments. Evidence from animal learning studies suggests that this operates via two distinct processes, a goal-directed component that involves learning of associations between responses and the incentive value of outcomes (response–outcome or stimulus–response–outcome learning), and a habit learning component that involves learning associations between stimuli (or context) and responses (stimulus–response learning) (Dickinson, 1985). Substantial neurobiological evidence supports the existence of distinct goal-directed and habit learning systems in rats, implicating the prefrontal cortex and dorsomedial striatum in the former and the dorsolateral striatum in the latter (Balleine and Dickinson, 1998a; Corbit and Balleine, 2003; Killcross and Coutureau, 2003; Yin et al., 2004, 2005). However, to our knowledge, no study has yet attempted to differentiate between the neural systems involved in goal-directed or habit learning in the human brain.

Received Nov. 21, 2006; revised March 5, 2007; accepted March 6, 2007.

This work was supported by grants from the Gimbel Discovery fund for Neuroscience and National Institute of Mental Health Grant R03MH075763 to J.O.D. We thank Steve Flaherty and Mike Tyszka for technical assistance and advice.

Correspondence should be addressed to John P. O'Doherty, Division of Humanities and Social Sciences, California Institute of Technology, 1200 East California Boulevard, Mail Code 228-77, Pasadena, CA 91125. E-mail: jdoherty@hss.caltech.edu.

DOI:10.1523/JNEUROSCI.0564-07.2007

 0270-6474/07/274019-08$15.00/0

The canonical assay for distinguishing between goal-directed and habitual processes is the outcome-devaluation paradigm (Dickinson, 1985). In the typical study, a hungry animal is trained to perform an instrumental response for a particular food outcome. After this training, the outcome is devalued by feeding the animal on the food to induce a state of specific satiety for that food (Balleine and Dickinson 1998b). The animal is then tested for its propensity to perform the instrumental action in extinction. If performance is mediated by response–outcome learning, and therefore goal-directed learning, responding should be reduced relative to a condition in which the outcome has not been devalued. In contrast, if the habitual stimulus–response process controls responding, performance should be impervious to outcome devaluation.

To determine the brain regions involved in goal-directed learning in the human brain, we trained subjects to perform two different instrumental actions to obtain two different food outcomes. We then devalued one of the two food outcomes by feeding subjects to satiety on that food, whereas the values of other foods not eaten remained high. This technique has been used in previous functional magnetic resonance imaging (fMRI) studies to determine brain regions involved in representing the value of olfactory or food stimuli, as well as in encoding the value of Pavlovian conditioned stimuli (O'Doherty et al., 2000; Gottfried et al., 2003; Gottfried and Dolan, 2004). However, this technique has never been used before to establish brain regions involved in implementing goal-directed instrumental action selection as op-

posed to Pavlovian conditioning that, in contrast, involves passive learning of stimulus-outcome associations. Here, after devaluation, we scanned subjects while they performed both actions in extinction, and tested for brain areas responding during action selection that showed sensitivity to the change in value of the associated outcomes, as such area(s) would be candidate regions for implementing goal-directed instrumental learning in humans. We focused in particular on the orbitofrontal cortex and amygdala, because these regions have been implicated previously in mediating instrumental outcome devaluation effects in nonhuman primates (Baxter et al., 2000; Izquierdo et al., 2004).

Materials and Methods

Subjects. Nineteen healthy right-handed individuals (eight females, 11 males; mean age, 28.8 ± 2.89; range, 18–66) participated in the experiment. An additional four subjects were scanned, but were excluded from any additional analysis, three because of the absence of learning (choosing the low-probability action >75% of the time), and one because of reported confusion at test. The subjects were preassessed to exclude those with a previous history of neurological or psychiatric illness. The eating attitudes test (EAT-26) (Garner et al., 1982) was administered and indicated no eating disorders in any of the subjects (mean score, 3.7 ± 0.92; range, 0–17; all scores were under the 20 point cutoff). Before participation in the experiment, the subjects were prescreened to ensure that they found tomato juice, chocolate milk, and orange juice to be pleasant, and to ensure that they were not overweight, on a diet, or planning to go on a diet.

Subjects were asked to fast for at least 6 h before their scheduled arrival time at the laboratory, but were permitted to drink water. All subjects gave informed consent and the study was approved by the Institutional Review Board of the California Institute of Technology.

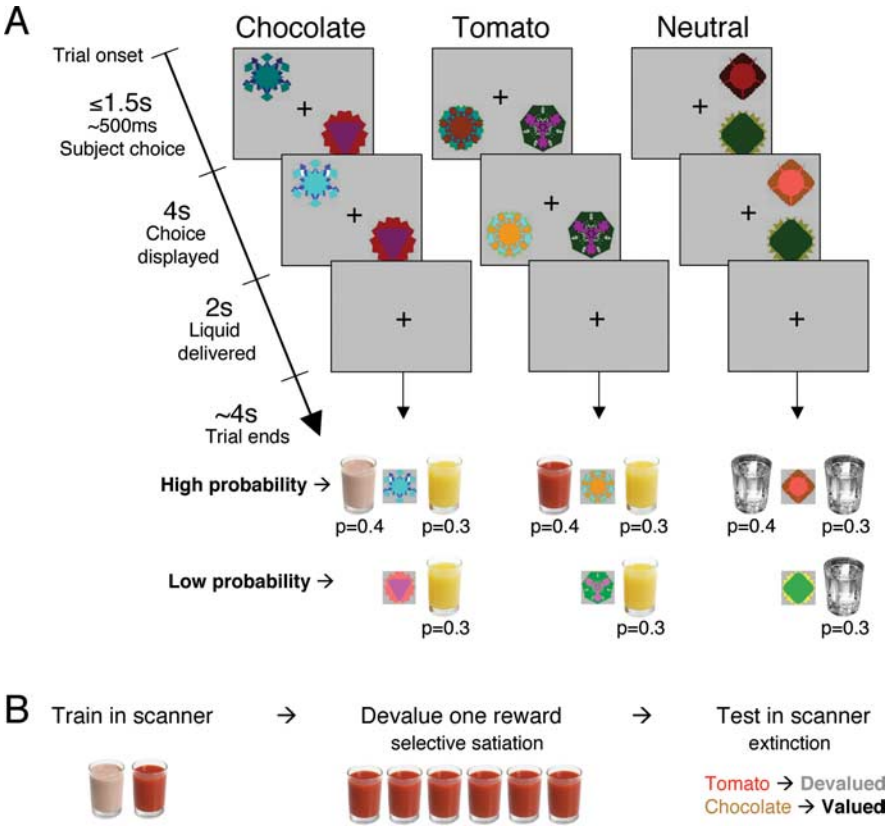

Figure 1. ***A***, Instrumental task illustration. The different actions available in the tomato, chocolate, and neutral conditions were signified by different arbitrary stimuli placed in one of four different locations. On each trial, subjects had to choose between two possible actions, one leading to a high probability of a food outcome ($p = 0.7$) and the other a low probability ($p = 0.3$). Depending on the condition, the high-probability action yielded tomato juice or chocolate milk with $p = 0.4$, a common outcome (orange juice) with $p = 0.3$, or else nothing. The low-probability action yielded the common orange juice outcome with $p = 0.3$. Once an action was chosen, the stimulus signifying that action was illuminated, and 4 s later the outcome was delivered. ***B***, Illustration of the experimental design. Subjects were trained in the scanner to choose two high-probability actions, one that led to chocolate milk, and one that led to tomato juice. They were then removed from the scanner and invited to consume either tomato (illustrated here) or chocolate to satiety, resulting in a selective decrease in the pleasantness of that food (selective satiation). They then underwent the same instrumental choice procedure in the scanner in extinction (the tomato or chocolate outcomes were no longer delivered, although the orange juice outcome continued to be delivered to maintain some degree of responding on both actions). At test, the condition involving the food eaten (in this example tomato) is designated "devalued" and the other is called "valued."

Stimuli. The three liquid-food rewards were chocolate milk (Nestle, Vevey, Switzerland), tomato juice (Campbell's, Camden, NJ), and orange juice (Sunny-Delight, Cincinnati, OH). The criteria for selecting these liquid-foods were to be administrable in liquid form, palatable at room temperature, and distinguishable in their flavor and texture to help facilitate sensory-specific satiety effects and minimize the likelihood of the subjects developing a generalized satiety to all liquid foods. In addition, we also used an effectively neutral control tasteless solution, which consisted of the main ionic components of human saliva (25 mM KCl and 2.5 mM $NaHCO_3$). The food rewards were delivered by means of separate electronic syringe pumps (one for each liquid) positioned in the scanner control room. These pumps transferred the food stimuli to the subject via ~10 m long polyethylene plastic tubes (6.4 mm diameter), the other end of which were held between the subject's lips like a straw while they lay supine in the scanner.

Task. The task consisted of three trial types: chocolate, tomato, or neutral, whose occurrence was fully randomized throughout the experiment (Fig. 1*A*). On each trial, subjects were faced with the choice between two possible actions. Each trial type had unique pairs of images representing those actions. On the chocolate and tomato trials, one action delivered the respective reward with a probability of $p = 0.4$. In addition, both actions delivered a common outcome of orange juice with an overall probability of $p = 0.3$ with the constraint that the orange juice and the chocolate or tomato rewards could not be delivered on the same trial. Therefore, the overall probability of a food outcome was $p = 0.7$ for the high-probability action, but $p = 0.3$ for the low-probability actions for each trial type. To provide a control condition against which to assess the effects of the rewards on neural activity, the neutral solution was delivered with the probabilities of $p = 0.7$ and $p = 0.3$ after the two actions on neutral trials.

Each action was uniquely signified by a specific arbitrary, effectively neutral fractal stimulus, which was presented in one of four locations on the screen: top left, top right, bottom left, or bottom right. Subjects could choose a given action by selecting one of four button presses on a response pad corresponding to each of the four locations on the screen. Subjects were trained to select all of the locations with ease by pressing one of the four buttons in a row with the following correspondence: 1, top left; 2, top right; 3, bottom left; 4, bottom right. Each stimulus–action pair was randomly assigned to one of the four positions at the beginning and remained constant throughout the experiment. A unique spatial location was assigned to the high-probability action in all three trial-type pairs. The specific assignment of arbitrary fractal stimuli and spatial position to each particular action was fully counterbalanced across subjects. The subjects' task on each trial was to choose one of the two possible

available actions. If a response was not registered before 1.5 s, a response omission was indicated to the subject, and the trial was aborted. When an action had been selected, the stimulus signaling that action increased in brightness and 4 s later the screen was cleared. Immediately after this, 0.75 ml of liquid food reward or neutral control solution was delivered or else no liquid stimulus was delivered (according to the reward schedule associated with the particular action chosen). This delivery was followed by an intertrial interval drawn from a Poisson distribution with a mean of 4 s.

Experimental design. The outcome-devaluation paradigm was the critical manipulation in this study, whereby the value of one outcome associated with a particular instrumental action is selectively decreased while the value of the other outcome associated with another particular instrumental action is maintained. The technique used to achieve this outcome manipulation was selective satiation, whereby the value of one food reward is decreased by feeding a subject to satiety on that food, while the value of other foods not eaten to satiety remain high. Subjects were required to come into the experiment hungry as a result of the 6 or more hours of fasting. Before starting the experimental task, we collected behavioral ratings, including hunger level (0, full; +10, starving) and pleasantness of the liquid foods (−5, very unpleasant; +5, very pleasant). Subjects underwent two 30 min scanning sessions, training and test, each consisting of 150 trials (50 trials per condition: chocolate, tomato, and neutral). There was a break in between the two sessions during which subjects were fed to satiety on one of the two foods outside of the scanner. All three trial types were pseudorandomly intermixed throughout both of the sessions. Before the experiment, subjects were told that there were three pairs of fractal patterns, and on each trial, one of these pairs would be displayed. They were instructed to select one of the possible actions on each trial. They were told that after their choices they could receive 0.75 ml of liquid food, the same quantity of a neutral tasteless solution, or nothing. They were not told which action was associated with which particular outcome, but they were told that one of each pair of actions was associated with a higher probability of obtaining an outcome than the other. For the first training session subjects were instructed to learn to choose the actions that led to high probabilities of pleasant liquid foods, including chocolate milk and tomato juice. Choosing this action led to a chance of obtaining chocolate milk ($p = 0.4$) or orange juice ($p = 0.3$) in the chocolate condition, and tomato juice ($p = 0.4$) or orange juice ($p = 0.3$) in the tomato condition. After the subjects had learned to preferentially choose the actions that gave them the best chance of obtaining a juice reward, they were then removed from the scanner and invited to eat either tomato soup (example indicated in Fig. 1*B*) or chocolate ice-cream to satiety (selective satiation), until they did not want to eat any more, and the pleasantness rating for that food had decreased (devaluation). The specific food used for devaluation (tomato or chocolate) was fully counterbalanced across subjects. This selective outcome devaluation procedure served to devalue one of the outcomes associated with a particular instrumental action, although leaving the value of the outcome associated with the other action intact. Behavioral ratings of hunger level and pleasantness of the liquid foods were collected before the next session of the experiment. To test the effects of the devaluation procedure, subjects were then placed back into the scanner to resume the task, in which they were invited to choose between the actions that led to different food outcomes at training. During the test session, they were presented with the same trial types involving the same actions and once again had to select whichever action they preferred. The chosen stimulus increased in brightness as it did during training; however, on this occasion, the chocolate or tomato outcomes were no longer presented (i.e., the subjects were tested in extinction for these outcomes). That is, the devalued and nondevalued outcomes were never presented again to the subjects during the test. To maintain some degree of responding on both actions (even the devalued one), we still presented the nondevalued orange juice outcome so that the overall outcome was now available with equal probability on the two available actions ($p = 0.3$ each), just as this orange juice outcome had been available with equal probability during training. The use of an extinction procedure ensured that the subjects only use information about the value of the outcome by making use of the previously learned associations between that outcome and a particular action, as otherwise, if the tomato and chocolate outcomes were presented again at test, subjects could relearn a new association, thereby confounding stimulus–response and response–outcome contributions. Moreover neural responses related to extinction per se are not a confound in this study, because both valued and devalued actions are presented in extinction and therefore, a direct comparison between the two actions controls for the overall effects of extinction (which will be present during selection of both actions).

Decreased responding to the action associated with the devalued outcome compared with the action associated with the valued outcome is the indication of goal-directed performance. After the test, behavioral ratings for hunger level and pleasantness of the liquid foods were again collected.

fMRI data acquisition. The functional imaging was conducted by using a Siemens (Erlangen, Germany) 3.0 Tesla Trio MRI scanner to acquire gradient echo T2*-weighted echo-planar images (EPIs) with BOLD (blood oxygenation level-dependent) contrast. To optimize functional sensitivity in the orbitofrontal cortex (OFC), we used a tilted acquisition in an oblique orientation of 30° to the anterior–posterior commissure line (Deichmann et al., 2003). In addition, we used an eight-channel phased array coil which yields a 40% signal increase in signal in the medial OFC over a standard head coil. Each volume comprised 32 axial slices. A total of 900 volumes (30 min) were collected during the experiment in an interleaved-ascending manner. The imaging parameters were as follows: echo time, 30 ms; field of view, 192 mm; in-plane resolution and slice thickness, 3 mm; repetition time, 2 s. Whole-brain high resolution T1-weighted structural scans ($1 \times 1 \times 1$ mm) were acquired from the 19 subjects and coregistered with their mean EPI images and averaged together to permit anatomical localization of the functional activations at the group level. Image analysis was performed using SPM2 (statistical parametric mapping software, Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). Temporal normalization was applied to the scans, each slice being centered to the middle of the scan. To correct for subject motion, the images were realigned to the first volume, spatially normalized to a standard T2* template with a resampled voxel size of 3 mm, and spatially smoothed using a Gaussian kernel with a full width at half maximum of 8 mm. Intensity normalization and high-pass temporal filtering (using a filter width of 128 s) were also applied to the data (Friston et al., 1994).

fMRI data analysis of testing session. The event-related fMRI data were analyzed by creating regressors composed of sets of δ (stick) functions at the time of action selection. Separate regressors were created for different actions to model activity at the time of the response to devalued high-probability actions (H-DEV), devalued low-probability actions (L-DEV), valued high-probability actions (H-VAL), valued low-probability actions (L-VAL), neutral high-probability actions (H-NEU), and neutral low-probability action s (L-NEU). Regressors for the receipt of the orange juice and neutral outcomes were also created. All of these regressors were convolved with a hemodynamic response function. In addition, the six scan-to-scan motion parameters produced during realignment were included to account for residual effects of movement. These regressors were then entered into a regression analysis against the fMRI data for each individual subject. Linear contrasts of regressor coefficients were computed at the single subject level to enable comparison between the H-VAL, L-VAL, H-DEV, L-DEV, H-NEU, and L-NEU actions. The results from each subject were taken to a random effects level by including the contrast images from each single subject into a one-way ANOVA with no mean term. The main contrast used to test for the effects of goal-directed learning is as follows: [H-VAL − L-VAL] − [H-DEV − L-DEV], which tests for those areas which respond differently to the high- compared with the low-probability action choices in devalued compared with nondevalued trials.

The structural T1 images were coregistered to the mean functional EPI images for each subject and normalized using the parameters derived from the EPI images. Anatomical localization was performed by overlaying the *t* maps on a normalized structural image averaged across subjects and with reference to an anatomical atlas.

Only activations in areas of a priori interest are featured in the results. Our a priori regions of interest (ROIs) are the orbital and medial pre-

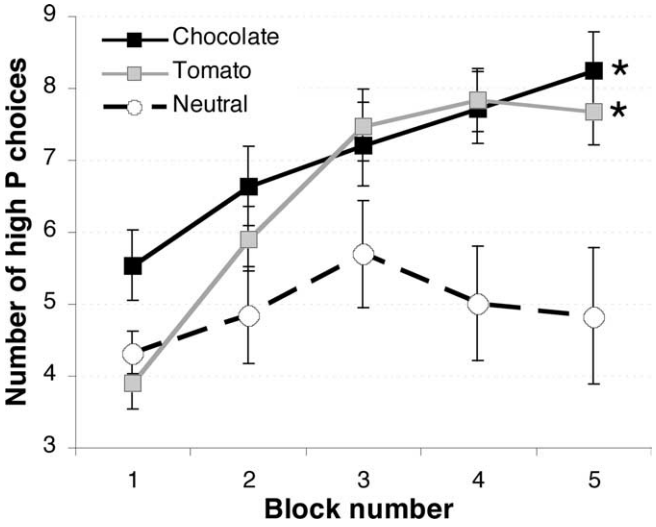

Figure 2. Learning curves. Total number of high-probability action choices over five 10-trial blocks shown averaged across 19 subjects during training. Over the course of training, subjects increasingly favored the high-probability actions associated with tomato juice or chocolate milk over their low-probability counterparts, but this was not the case for the neutral condition where subjects were indifferent between the high- and low-probability actions (*$p < 0.0005$, one-tailed). Error bars indicate SEM.

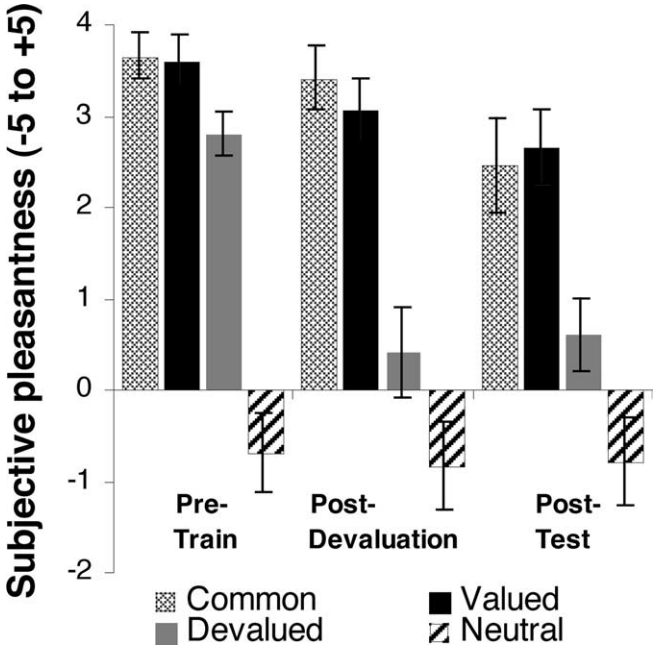

Figure 3. Subjective pleasantness ratings on a scale of −5 (very unpleasant) to +5 (very pleasant) before training, after devaluation, and after test. The rating for the food eaten (devalued) significantly decreased compared with the food not eaten (valued) after the selective devaluation procedure (interaction at $p < 0.01$). Error bars indicate SEM.

frontal cortex, dorsolateral prefrontal cortex, anterior cingulate cortex, ventral and dorsal striatum, and amygdala, because these areas have been implicated previously in reward-related processing and learning (O'Doherty et al., 2004).

For the time course plots, we located functional ROIs within an individual subject's medial and central OFC and extracted event-related responses from the peak voxel for that subject. These single-subject time courses were then averaged across subjects.

In addition to the analysis of the test session data, we also conducted an analysis of goal-directed learning during the training session by contrasting action choices associated with reward (tomato and chocolate) and action choices associated with neutral outcomes. We performed a small-volume correction using the resulting image from the main contrast at test.

Results

Behavioral results

Effects of training during instrumental conditioning

Figure 2 shows learning curves for the high-probability actions associated with tomato juice and chocolate milk outcomes over the course of training. In the last 10-trial block of training, subjects chose the high-probability action significantly more often than the low-probability action in both the chocolate ($t_{(18)} = 6.29$; $p < 0.0005$, one-tailed) and the tomato ($t_{(18)} = 6.06$; $p < 0.0005$, one-tailed) conditions. This indicates that subjects learned to choose the instrumental action associated with the most food reward in both conditions. On the contrary, subjects did not learn to choose the high-probability action more than the low-probability action in the last block of the neutral condition ($t_{(18)} = -0.25$; $p = 0.809$, two-tailed), indicating that subjects were indifferent as to whether they obtained the effectively neutral control solution. In addition, we found no significant difference in the number of high-probability action choices made in the tomato and chocolate condition in the last 10 trials, indicating that the instrumental actions associated with these two food rewards were learned equally well ($t_{(18)} = 1.58$; $p = 0.132$, two-tailed).

Effects of devaluation procedure on the subjective value of the food outcomes

Subjects showed a significant reduction in subjective hunger ratings after the selective satiation procedure ($t_{(18)} = 11.59$; $p < 0.0005$, one-tailed). Mean hunger ratings were 7.34 (±0.25 SEM) before satiety, but dropped to 1.79 (±0.41 SEM) after satiety. Subjective pleasantness ratings for the three different food rewards before and after feeding to satiety with one of the foods are plotted in Figure 3. Consistent with specific satiation, the subjective pleasantness of the food eaten (devalued) decreased markedly, whereas the pleasantness of the foods not eaten did not show any such decrease. This effect was statistically significant as shown by a significant interaction effect in a repeated-measures two-way ANOVA with one factor food type (valued vs devalued) and the other factor time (before and after feeding; $F_{(1,18)} = 8.63$; $p < 0.001$).

Effects of devaluation procedure on instrumental responding during test

Figure 4*A* shows that during the test, the valued high-probability response was performed more frequently than the devalued response. However, this devaluation effect was most prominent during the first 10-trial block before the participants had the opportunity to learn that the chocolate and tomato outcomes were no longer presented, which led to a convergence of the valued and devalued actions across the test. Consequently, to assess the devaluation effect, we compared the performance on the first trial block of testing with that on the last block of training (Fig. 4*B*). A two-way ANOVA with one factor condition (valued vs devalued) and the other factor session (training vs test), revealed a significant condition by session interaction ($F_{(1,18)} = 9.642$; $p < 0.01$). Expressing the number of high-probability responses in the first 10 trials of test as a percentage of the number of high-probability responses in the last 10 trials of training also yields a significant difference between valued (87.7%) and devalued (55.3%) conditions ($t_{(18)} = -3.04$; $p < 0.005$, one-tailed).

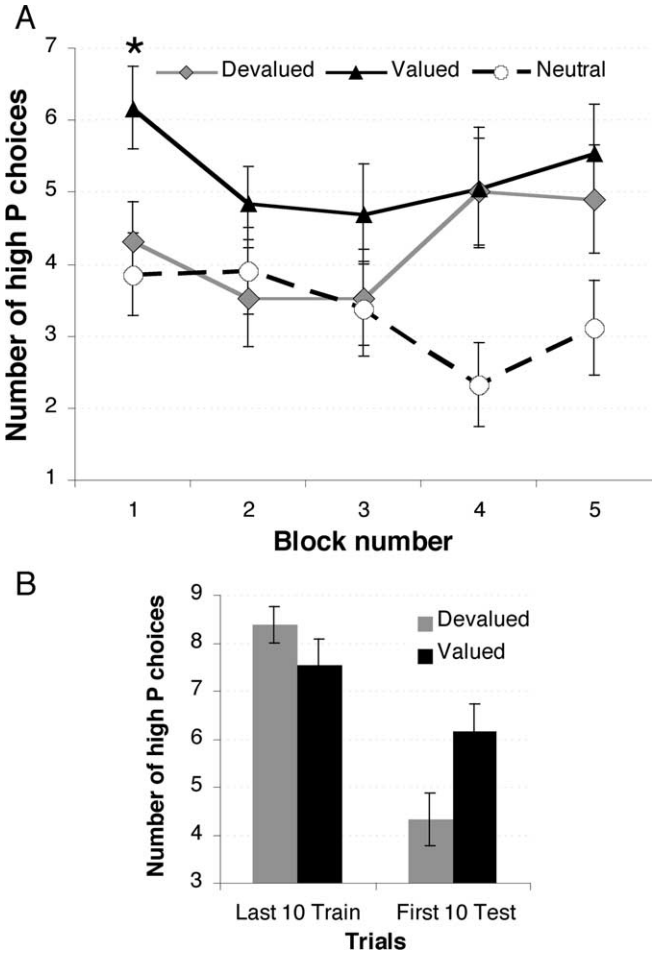

Figure 4. ***A***, Extinction curves. Total number of high-probability action choices over five 10-trial blocks averaged across subjects during extinction testing. The number of choices of the high-probability action in the first block was significantly greater in the valued compared with the devalued condition, indicating that subjects modulated their instrumental responses as a function of the change in value of the associated food outcomes (*$p < 0.05$, one-tailed). ***B***, After devaluation, subjects reduced their choices of the high-probability action associated with the devalued food significantly more than that of the valued food (interaction with $p < 0.01$). Error bars indicate SEM.

More specifically, after devaluation, subjects no longer favored choice of the high-probability action in the devalued condition ($t_{(18)} = -1.25$; $p = 0.227$, two-tailed), whereas, subjects still favored choice of the valued action ($t_{(18)} = 2.02$; $p < 0.05$, one-tailed). Moreover, the devalued action was chosen less frequently than the valued one ($t_{(18)} = 1.93$; $p < 0.05$, one-tailed). These results show that subjects were able to modulate their instrumental responses as a function of a change in the value of the associated outcome, thereby providing direct behavioral evidence of goal-directed learning in our paradigm.

Effects of devaluation on reaction times

We analyzed the effects of the devaluation procedure on reaction times (RT) for the different actions during the 10 trials of test using a two-way ANOVA with one factor condition (valued vs devalued) and the other factor action type (high vs low probability). We found no significant effect of condition and no significant interaction between condition and action type on RTs during test, suggesting that RTs were not modulated as a function of the devaluation procedure.

Neuroimaging results

Identifying the neural correlates of goal-directed learning

To identify brain regions involved in mediating the goal-directed component of instrumental conditioning, we looked for a significant condition (valued vs devalued) by action (high vs low probability) interaction during the test period (in extinction). This analysis revealed significant effects in the medial OFC ($x = 0$, $y = 33$, $z = -24$ mm; $Z = 3.33$; $p < 0.001$) (Figure 5*A*). Activity in this region survived correction for small volume at $p < 0.01$ with an 8 mm sphere centered on coordinates derived from previous studies that reported medial OFC responses during instrumental conditioning [$x = 0$, $y = 33$, $z = -18$ mm and $x = -6$, $y = 30$, $z = -21$ mm from Kim and O'Doherty (2006); $x = 3$, $y = 30$, $z = -21$ mm from Daw and O'Doherty et al. (2006)]. Other areas showing significant effects in this contrast were the right central OFC ($x = 24$, $y = 45$, $z = -6$ mm; $Z = 3.23$; $p < 0.001$) (Figure 5*C*) and an area in the left lateral OFC bordering on the inferior prefrontal cortex [$x = -39$, $y = 30$, $z = -15$ mm; $Z = 3.11$; $p < 0.001$; $p < 0.05$ when corrected for small volume using a sphere of 8 mm centered on the coordinates $x = -36$, $y = 27$, $z = -21$ mm (from Kim and O'Doherty, 2006)]. We also extracted trial averaged time-course data from the peak voxels in the medial and central OFC from each subject and then averaged across subjects (Fig. 5*B*,*D*). These areas showed an increase in activity when subjects chose the high-probability action in the valued condition, but a decrease in activity when subjects chose the high-probability action in the devalued condition, suggesting that these regions are sensitive to the incentive value of the associated outcomes associated with particular instrumental actions (even in extinction). Moreover, these areas showed an increase in activity on trials in which subjects chose the low-probability action in the devalued condition, suggesting that the incentive value of the alternative action that was not associated with the devalued outcome was increased as a result of the devaluation procedure.

To further exclude the possibility that activity in these areas was specific to the extinction context during the test phase, we examined responses during action selection of the rewarding outcomes during the last 20 trials of the learning phase, once subjects had learned the action–outcome associations, but before extinction occurred. This analysis revealed significant effects in the medial OFC (Fig. 6) that survived correction for small volume [family-wise error (FWE)-corrected, $p < 0.05$] within an ROI defined by areas showing significant effects in the main interaction contrast of goal-directed learning during test at $p < 0.001$, described above. These results indicate that activity in the medial OFC is not specific to the extinction context during test, but rather is involved in goal-directed instrumental action selection more generally.

Testing for effects of habit learning

Next, we looked for regions that did not show modulation in their activity as a function of devaluation of the associated instrumental outcomes, to address the possibility that, although habits are not controlling behavior in the present paradigm (because of the limited training used), some brain regions may still manifest neural responses consistent with habitization (as would be consistent with a gating hypothesis of goal-directed vs habit-learning) (Killcross and Coutureau, 2003; Daw et al., 2005). To do this, we performed a conjunction analysis, testing for regions that showed significant effects during test on trials involving choice of the valued and on trials involving choice of the devalued high actions compared with the action choices in the neutral condition.

Regions showing similar response profiles to actions associated with the valued and devalued actions in extinction would be candidate areas for mediating the habit component of instrumental conditioning. We found no significant effects at $p < 0.001$ in any of our regions of interest, consistent with habit learning, although weak effects bordering significance at $p < 0.001$ were found in a region of far posterior caudate (tail of caudate) suggestive of a possible contribution of this area to the habit learning component ($x = -27$, $y = -36$, $z = 12$ mm; $Z = 3.08$). Because the effects in this region did not quite reach our criteria for significance, we refrain from drawing strong conclusions. Nevertheless, this area may warrant additional study as a possible contributor to habit learning processes in humans.

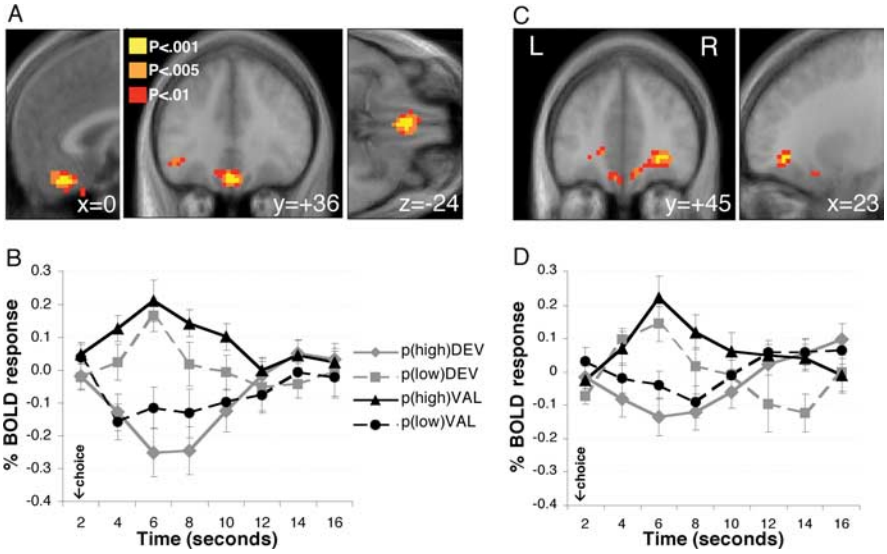

Figure 5. Neural correlates of goal-directed learning, as revealed by an interaction contrast between condition [valued (VAL) vs devalued (DEV)] and action choice (high vs low probability) performed on the fMRI data obtained during test. Voxels significant at $p < 0.001$ (uncorrected) are shown in yellow. To show extent of activation, we also show $p < 0.005$ in orange and $p < 0.01$ in red. The coordinate values for each section shown are provided on the bottom right of each image. ***A***, A region of the medial OFC showing a significant modulation in its activity during instrumental action selection as a function of the value of the associated outcome (mOFC; $x = -3, y = 36, z = -24$ mm; $Z = 3.29$; $p < 0.001$). ***B***, Time-course plots derived from the peak voxel (from each individual subject) in the mOFC during trials in which subjects chose each one of the four different actions (choice of the high- vs low-probability action in either the valued or devalued conditions). ***C***, A region of the right central OFC also showing a significant interaction effect (lOFC; $x = 24, y = 45, z = -6$ mm; $Z = 3.19$; $p < 0.001$). ***D***, Time course plots from the peak voxel (from each individual subject) in the right lOFC. Error bars indicate SEM.

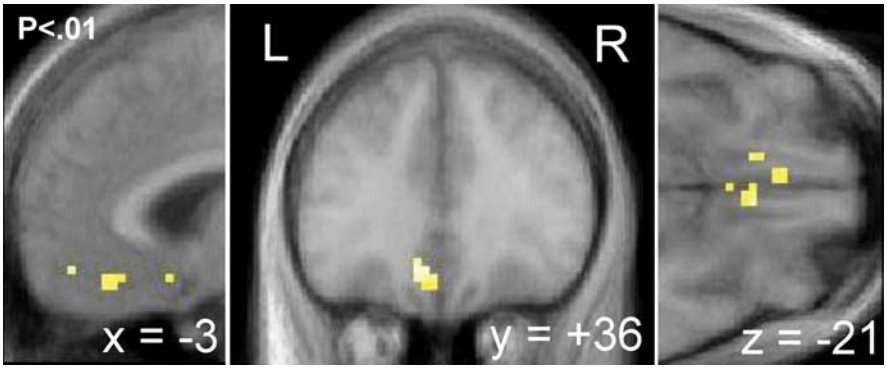

Figure 6. Neural correlates of goal-directed learning, as revealed by a contrast between action choices associated with rewarding versus neutral outcomes during training. Voxels significant at $p < 0.01$ (uncorrected) are shown, centered on the region in medial OFC ($x = -3, y = 36, z = -21$ mm; $Z = 2.44$; $p < 0.01$] that survived correction for small volume (FWE-corrected $p < 0.05$) using the image obtained from the interaction analysis during test at $p < 0.001$.

Discussion

Here, we provide evidence with fMRI that both medial and lateral regions of the orbitofrontal cortex show neural responses during performance of instrumental actions that reflect the incentive value of an associated outcome. Critically, these effects were demonstrated using a reinforcer devaluation procedure that was tested in extinction, which allows us to disambiguate goal-directed response–outcome (or stimulus–response–outcome) learning processes from stimulus-response learning. The finding that responses in the orbitofrontal cortex are sensitive to the incentive value of instrumental actions indicates that this region is likely to be involved in the goal-directed component of instrumental learning.

Our finding of a prefrontal locus for goal-directed learning in humans resonates with similar findings implicating the prefrontal cortex in this function in rats (Balleine and Dickinson, 1998a; Killcross and Coutureau, 2003; Ostlund and Balleine, 2005). Our present findings are also consistent with previous reports of a role for the ventromedial prefrontal cortex in behavioral choice in humans, as evidenced by significant responses in orbitofrontal cortex related to subsequent behavioral choice during reversal learning (O'Doherty et al., 2003), and reports of expected value signals in these areas during performance of instrumental choice tasks (Tanaka et al., 2004; Daw et al., 2006; Kim et al., 2006). Furthermore, single neurons in orbitofrontal cortex have been found to flexibly encode the value of expected outcomes during instrumental reversal learning in both rats and nonhuman primates (Thorpe et al., 1983; Schoenbaum et al., 2003; Schoenbaum and Roesch, 2005) whereas neurons in medial prefrontal cortex have been found to encode response–reward associations (Matsumoto et al., 2003). The results shown here are also compatible with lesion studies in nonhuman primates that indicate that orbitofrontal cortex-lesioned animals are unable to flexibly adapt instrumental responding after reinforcer devaluation (Baxter et al., 2000; Izquierdo et al., 2004).

An alternative account of our results is that modulation of activity in orbitofrontal cortex as a function of reinforcer devaluation is driven by stimulus–outcome and not response–outcome or stimulus–response–outcome associations. Gottfried et al. (2003) showed that a region of the midcentral OFC tracked changes in the value of a pavlovian conditioned stimulus after devaluation of associated food odors. O'Doherty et al. (2002) reported a region of central OFC responding during expectation of a pleasant taste stimulus, as signaled by the previous presentation of an arbitrary conditioned stimulus. These findings certainly implicate the OFC in stimulus–outcome learning even in

the absence of instrumental choice behavior. However, other studies clearly show that responses in orbitofrontal cortex are sensitive to instrumental contingencies. O'Doherty et al. (2003) showed that activity in this region is significantly enhanced when subjects are performing instrumental actions to obtain reward as opposed to passively receiving rewards (thereby involving only stimulus–outcome associations). Furthermore, Arana et al. (2003) demonstrated increased activity in medial orbitofrontal cortex during a menu preference task in which subjects had to choose specific high incentive menu items from the menu as opposed to passively viewing these items. Moreover, lesions of orbitofrontal cortex produce robust impairments on instrumental choice tasks in both humans and nonhuman primates (Bechara et al., 1994, 2000; Rolls et al., 1994; Hornak et al. 2003; Fellows and Farah, 2003). In addition, a recent single-unit recording study in rats found evidence for the encoding of response-related information in OFC neurons, whereby some neurons responded selectively according to the direction in which the rat moved to attain reward (Feierstein et al., 2006). Response-related value encoding was also clearly found in rat OFC in a study by Roesch et al. (2006). These studies suggest that the orbitofrontal cortex contributes to instrumental as well as pavlovian learning processes.

It is interesting to note that in a previous pavlovian devaluation study by Gottfried et al. (2003), modulatory effects of reinforcer devaluation were found in central, but not medial OFC areas, whereas in the present study we found significant effects of instrumental devaluation in both central, lateral, and medial areas. This raises the possibility that the medial OFC may be more involved in the goal-directed component of instrumental conditioning whereas the central OFC may be more involved in pavlovian stimulus–outcome learning (as this area was found in both the present study and in the previous pavlovian devaluation study). This speculation is consistent with the known anatomical connectivity of these areas in which lateral and central areas of OFC (Brodmann areas 12/47, 11, and 13) receive input primarily from sensory areas, consistent with a role for these areas in stimulus–stimulus learning, whereas the medial OFC (areas 14 and 25) receives input primarily from structures on the adjacent medial wall of prefrontal cortex such as cingulate cortex, an area often implicated in response selection and/or reward-based action choice (Carmichael and Price, 1996; Walton et al., 2004). It is also notable that although the majority of single-unit studies in monkeys have reported stimulus-related activity and not response-related selectivity in the OFC (e.g., Thorpe et al., 1983; Tremblay and Schultz, 1999), the majority of these studies have tended to record from lateral and central areas of the OFC (Brodmann areas 12/47 and 13, respectively), and not from more medial areas, with the possible exception of Padoa-Schioppa and Assad (2006). The direct homologues of human medial and central OFC in rats are also not clear. It is plausible that the more medial sectors of the OFC in humans correspond to regions considered part of medial prefrontal cortex in rats that have been more conclusively linked to goal-directed learning in rat lesion studies. Additional research will be needed to reach more definitive conclusions about the distinct contributions of subregions of the human orbitofrontal cortex to pavlovian and instrumental learning. Nevertheless, the results of the present study, when taken together with previous findings, strongly suggest that the orbitofrontal cortex (especially its medial aspect) plays a role in the goal-directed component of instrumental choice.

The "actor" in the actor/critic model uses afferent prediction errors to modify stimulus–response associations and could therefore correspond to the habit learning component of instrumental conditioning (Dayan and Balleine, 2002; Daw et al., 2005). However, when probing directly for habit learning signals during test, we did not find strong evidence of such signals in the dorsal striatum or elsewhere, apart from a weak and inconclusive effect in a far posterior region of dorsal striatum (corresponding to the tail of the caudate nucleus). The absence of clear habit-learning signals during the test could be accounted for by the fact that, behaviorally, subjects exhibited goal-directed and not habitual responding. This finding accords with previous studies of behavioral habitization. Not only is overtraining usually required for resistance to outcome revaluation (Adams, 1982; Dickinson et al., 1995; Holland, 2004), but when the actions are trained in the context of a choice between different outcomes, as in the present experiment, performance remains goal-directed even in the case of extensive overtraining (Colwill and Rescorla, 1985; Colwill and Rescorla, 1988).

Even so, the failure to find strong evidence of habit signals could have important implications for understanding how the goal and habit systems may interact during conditioning. One possibility is that neural circuits involved in implementing both forms of learning are always engaged during instrumental responding, regardless of which system is currently controlling behavior. Alternatively, the system controlling behavior at a particular time may dominate (in activity), while the other system remains silent. Our findings lend some support toward the latter possibility, suggesting that brain regions involved in habit learning are not manifest at least to the extent that they can be reliably detected with fMRI at the point when behavior is being controlled by the goal-directed system. The goal of the present study was to determine brain regions involved in implementing goal-directed learning and not to directly address brain regions involved in habit learning. As a consequence, subjects were exposed to only moderate training, a manipulation that successfully produced goal-directed learning in our subjects. Given that habit learning is suggested to control behavior after an instrumental action has been performed extensively, an important direction for future research will be to train subjects extensively on a given action before scanning in an attempt to induce habitization at the behavioral level.

To conclude, in the present study we have implicated the orbitofrontal cortex, particularly its medial aspect in mediating goal-directed instrumental learning in humans. By using a selective devaluation procedure in extinction, we have been able to provide the first direct evidence of response–outcome learning in the human brain, by successfully disambiguating neural activity related to the goal-directed component of instrumental conditioning from that pertaining to habitual or stimulus–response learning. Although we were able to uncover neural correlates for goal-directed learning, we did not find compelling evidence for regions involved in habit learning. Because subjects were exposed to only moderate training in the present study, these findings raise the possibility that the habit learning system may become more engaged after extensive training, when behavior is more directly under the control of the habit system. More generally, this study highlights the utility of using constructs and experimental methodologies derived from animal learning theory to gain insight into the neural mechanisms underlying human behavior.

References

Adams CD (1982) Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q J Exp Psychol 34B:77–98.

Arana FS, Parkinson JA, Hinton E, Holland AJ, Owen AM, Roberts AC (2003) Dissociable contributions of the human amygdala and orbitofrontal cortex to incentive motivation and goal selection. J Neurosci 23:9632–9638.

Balleine BW, Dickinson A (1998a) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacol 37:407–419.

Balleine BW, Dickinson A (1998b) The role of incentive learning in instrumental outcome revaluation by sensory-specific satiety. Anim Learn Behav 26:46–59.

Baxter MG, Parker A, Lindner CC, Izquierdo AD, Murray EA (2000) Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. J Neurosci 20:4311–4319.

Bechara A, Damasio AR, Damasio H, Anderson SW (1994) Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 50:7–15.

Bechara A, Damasio H, Damasio AR (2000) Emotion, decision making, and the orbitofrontal cortex. Cereb Cortex 10:295–307.

Carmichael ST, Price JL (1996) . Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. J Comp Neurol 371:179–207.

Colwill RM, Rescorla RA (1985) Instrumental responding remains sensitive to reinforcer devaluation after extensive training. J Exp Psychol Anim Behav Processes 11:520–536.

Colwill RM, Rescorla RA (1988) The role of response-reinforcer associations increases throughout extended instrumental training. Anim Learn Behav 16:105–111.

Corbit LH, Balleine BW (2003) The role of prelimbic cortex in instrumental conditioning. Behav Brain Res 146:145–157.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neurosci 8:1704–1711.

Daw ND, O'Doherty JP, Dayan P, Seymour P, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36:285–298.

Deichmann R, Gottfried JA, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. NeuroImage 19:430–441.

Dickinson A (1985) Actions and habits: the development of a behavioural autonomy. Philos Trans R Soc Lond B Biol Sci 308:67–78.

Dickinson A, Balleine B, Watt A, Gonzalez F, Boakes RA (1995) Motivational control after extended instrumental training. Anim Learn Behav 23:197–206.

Feierstein CE, Quirk MC, Uchida N, Sosulski DL, Mainen ZF (2006) Representation of spatial goals in rat orbitofrontal cortex. Neuron 51:495–507.

Fellows LK, Farah MJ (2003) Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. Brain 126:1830–1837.

Friston KJ, Tononi G, Reeke Jr GN, Sporns O, Edelman GM (1994) Valuedependent selection in the brain: simulation in a synthetic neural model. Neuroscience 59:229–243.

Garner DM, Olmsted MP, Bohr Y, Garfinkel PE (1982) The eating attitudes test: psychometric features and clinical correlates. Psychol Med 12:871–878.

Gottfried JA, Dolan RJ (2004) Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. Nat Neurosci 7:1144–1152.

Gottfried JA, O'Doherty JP, Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. Science 301:1104–1107.

Holland PC (2004) Relations between pavlovian-instrumental transfer and reinforcer devaluation. J Exp Psychol Anim Behav Processes 30:104–117.

Hornak J, Bramham J, Rolls ET, Morris RG, O'Doherty JP, Bullock PR, Polkey CE (2003) Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. Brain 126:1691–1712.

Izquierdo A, Suda RK, Murray E (2004) Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. J Neurosci 24:7540–7548.

Killcross AS, Coutureau E (2003) Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb Cortex 13:400–408.

Kim H, Shimojo S, O'Doherty JP (2006) Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. PLoS Biol 4:e233.

Matsumoto K, Suzuki W, Tanaka K (2003) Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science 301:229–232.

O'Doherty J, Critchley H, Deichmann R, Dolan RJ (2003) Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. J Neurosci 23:7931–7939.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.

O'Doherty JP, Rolls ET, Francis S, Bowtell R, McGlone F, Kobal G, Renner B, Ahne G (2000) Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex. NeuroReport 11:893–897.

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ (2002) Neural responses during anticipation of a primary taste reward. Neuron 33:815–826.

Ostlund SB, Balleine BW (2005) Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. J Neurosci 25:7763–7770.

Padoa-Schioppa C, Assad JA (2006) . Neurons in the orbitofrontal cortex encode economic value. Nature 441:223–226.

Roesch MR, Taylor AR, Schoenbaum G (2006) Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. Neuron 51:509–520.

Rolls ET, Hornak J, Wade D, McGrath J (1994) Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. J Neurol Neurosurg Psychiatry 57:1518–1524.

Schoenbaum G, Roesch M (2005) Orbitofrontal cortex, associative learning, and expectancies. Neuron 47:633–636.

Schoenbaum G, Setlow B, Saddoris MP, Gallagher M (2003) Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. Neuron 39:855–867.

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neurosci 7:887–893.

Thorpe SJ, Rolls ET, Maddison S (1983) The orbitofrontal cortex: neuronal activity in the behaving monkey. Exp Brain Res 49:93–115.

Tremblay L, Schultz W (1999) . Relative reward preference in primate orbitofrontal cortex. Nature 398:704–708.

Walton ME, Devlin JT, Rushworth MF (2004) . Interactions between decision making and performance monitoring within prefrontal cortex. Nat Neurosci 7:1259–1265.

Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci 19:181–189.

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22:513–523.